Heavy Tails of Distributions of Words in Literary Texts

Adam J. Callahan & Gary E. Davis

Department of Mathematics, University of Massachusetts Dartmouth, 285 Old Westport Road, North Dartmouth, MA 02747, USA

Abstract

We examine a range of literary texts, of different genres, in English and other languages, and show that the type-token ratio – the ratio of the number of new words in an initial segment of the text to the number of words in that segment – is approximated as a regularly varying function. A careful examination of the type-token ratio of the regularly varying functions shows that, in all cases considered, there is an initial transition segment of text where the type-token ratio is well-approximated by the ratio of a logarithmic function and a power function. After that transition segment, the type-token ration is well-approximated by a pure power function. The existence of this segment may indicate a transition to a state of organized complexity. However, we observe that in some cases the type-token ratio is better approximated overall by a logarithmic decay, and in the absence of a model of word production and word use it is difficult to choose between these two competing descriptions. We also examine the Shannon entropy of initial segments of a text, from the word frequencies, and show that in all texts we examined the entropy increases logarithmically with the number of words. We apply this to the Voynich manuscript to show an unusual decrease of entropy in part of that text.

1 Introduction

The distribution of word frequencies in text has been studied extensively: see, for example, [1]-[13]. Recently Goncalves and Goncalves [13] showed that across a range of authors and publications of those authors, the ratio of the number of types (number of different words) to the number of tokens (number of words) varies approximately as a power law with the number of tokens. In this paper we consider the relationship between several variables, including the type-token ratio and the number of tokens, *within* a given text. Within given texts, whether several thousand words or several hundreds of thousands of words in length, whether in English or other languages, and for texts of different genres, we find that the type-token ratio - the ratio of new words to total words used, for the early part of the text - typically one or two thousand words, does *not* follow a power law that is typical for the entire text. Rather, we find that for some index $0 < d < 1$, the ratio of new word types in the first n words of the text to n is a function of the form $\frac{L(n)}{n^d}$ where L is a function that increases rapidly for small n and then changes only marginally after that. We characterize these functions as slowly varying with a variance that decays as a power function. The power law decay of the type-token ratio after an initial block of text produces an approximate power law decay for large runs of previously used words in the text.

We find that, unlike the running average of new word types, the running variance does not decay as a power law. Rather there are several classes of behavior for the decay of the running variance of new word types, including linear, logarithmic, piecewise linear, and piecewise linear-logarithmic.

Apart from making an appearance as new words, words in a text have a frequency of occurrence both in the complete text, and in the text to a given point. Focusing on the frequency of occurrence of a word in an initial segment of text, and taking that frequency as an estimate of probability of occurrence of the word to that point, we compute the empirical running Shannon entropy of the text. We find empirically that the running entropy increases approximately logarithmically with the length of text. We show that new, and infrequently used, words necessarily increase the entropy of the text to that point. The entropy increase is slowed by the ever-increasing use of previously used words in the text. Words that have, to a given point, been used relatively frequently, will produce a local entropy decrease, but these local decreases are swamped by the overall logarithmic increase of the entropy. Instances in text where the entropy decreases significantly for a period correspond therefore to sustained use of previously used words.

In what follows, text words were obtained by deleting all punctuation from a given text. In the process we transformed a text occurrence such as "I've" into a text word "Ive", and a hyphenated text occurrence such as "long-term" into a text word "longterm". This process can lead to some minor miscounting such as confounding "its" (possessive) with "its" coming from "it's".

2 Type-token ratio

2.1 Slowly varying functions

For a given text the type-token ratio is $\rho(n) := \frac{types(n)}{n}$, where $types(n)$ is the number of word types in the first n words of the text. The type-token ratio $\rho(n)$ is just the running average of the number of new words in an initial text segment of length n.

A real-valued function f of a positive integer variable n is *regularly varying* if for all $n \geq 1$, $\frac{f(mn)}{f(m)} \to n^k$as $m \uparrow \infty$ for some index k (ref. [14]). When the index $k = 0$ the regularly varying function f is *slowly varying.* Clearly, regularly varying functions are exactly those of the form $n^k L(n)$ where L is slowly varying, so regularly varying functions behave asymptotically as power functions. The function $log(n)$, for example, is slowly varying, and any function L for which $\lim_{n\to\infty} L(n)$ is positive and finite is slowly varying; in particular, constant functions are slowly varying. The variance $var(n) := Variance\{log(1), \ldots, log(n)\}$ of $log(n)$, for $n \geq 2$ is also a slowly varying function: it is an increasing function with $\lim_{n\to\infty} = 1$ (a consequence of Sterling's approximation to $n!$). By way of contrast, we will be interested in slowly varying functions L for which $var(L;n) := Variance\{L(1), L(2), \ldots, L(n)\}$ is itself slowly varying with negative index - we call such functions *ultra-slowly varying*. These are, roughly speaking, functions whose variance eventually decreases to 0 as a power function. The variance $var(L;n)$, rather than the variation $\sum_{i=1}^{n} |L(i+1) - L(i)|$, plays a critical role in distinguishing the slowly varying functions we see in text data from slowly varying functions such as log. We provide evidence that, for a wide variety of literary texts of different lengths, genres, and languages, the type-token ratio $\rho(n)$ is a regularly varying function of n: $\rho(n) = \frac{L(n)}{n^d}$ where $0 < d < 1$ and L is an ultra-slowly varying function. If L were constant this would yield an exact power law for the type-token ratio: in practice, we find that L is a non-constant, ultra-slowly varying function.

2.2 Power laws

To a first approximation, the type-token ratio is a power function of the number of words: $\rho(n) \approx \frac{A}{n^d}$. A linear least squares fit to a plot of $log(\rho(n))$ versus $log(n)$ yields the following estimates for the constants A and d for a variety of texts:

Power law decay $\frac{A}{n^d}$ of the running average of new words, for a variety of texts

Author - Text - Language	**# words**	**A**	**d**	r^2
Kant, I. - Kritik der reinen Vernunft - German	174877	13.059	0.459	0.987
Darwin, C. - Origin of Species - English	149133	12.254	0.458	0.984
Milton, J. - Paradise Lost - English	80007	6.505	0.3423	0.968
Lawrence, D.H. - England My England - English	64108	3.807	0.316	0.986
Balzac - Contes Bruns - French	60775	3.680	0.279	0.981
Wodehouse. P.G. - Leave It To Jeeves - English	52832	4.735	0.340	0.989
Dante Alighieri - Inferno - Italian	32265	2.923	0.250	0.985
Voynich manuscript	31851	2.536	0.196	0.959
Traditional - Beowulf - Old English	23732	5.487	0.352	0.956
Wittgenstein, L. - Tractatus - English	22580	4.600	0.385	0.991
Aaronsohn, A. - With the Turks in Palestine - English	17537	3.150	0.270	0.964
Einstein, A. - Sidelights on Relativity - English translation	8472	2.962	0.299	0.971
Greer, G. - Worlds Apart	3557			
Elliot, T.S.- Waste Land - English	3109	2.142	0.211	0.950
McKenzie, N. - The master networker (news item) - English	2266	1.898	0.200	0.971
Traditional - Hinemoa - Maori	2005	2.859	0.363	0.970

Table 1. The type-token ratio $\rho(n)$ is approximated by a power function of n over a variety of texts, authors, languages, and genres.

2.3 Power law or logarithmic decay?

Generally speaking, a researcher is very happy to discover a power law, or something closely approximating a power law, for a phenomenon in which they are interested. A major reason is that power laws indicate scale-invariant phenomena. Power laws are also indicative of a transition to self-organized critical behavior of a system [17]. The type-token ratios of texts we examined are well described by power laws, or by regularly varying functions $\frac{L(n)}{n^d}$ where $L(n)$ is ultra-slowly varying. However, the type-token ratio $\rho(n)$ is, in some instances, better described as a decreasing logarithmic function of n, while in other instances, a power law decay is a better fit to the data:

Logarithmic decay $A - Blog(n)$of the running average of new words

Author - Text	**A**	**B**	r^2
Milton, J. - Paradise Lost	1.1254	0.0894	0.992 (cf 0.968 for power law)
Lawrence, D.H. - England My England	0.8812	0.0709	0.942 (cf 0.986 for power law)
Wodehouse. P.G. - Leave It To Jeeves	0.8964	0.0722	0.954 (cf 0.989 for power law)
Voynich manuscript	1.2868	0.0931	0.9884 (cf 0.959 for power law)
Einstein, A. - Sidelights on Relativity	1.1200	0.1043	0.991 (cf 0.971 for power law)

Table 2. Comparison of power law and logarithmic fits to $\rho(n)$ as a function of n for selected texts.

From the perspective of the empirical data it is difficult to decide between a power law or logarithmic decay of the type-token ratio. What is lacking is a theoretical model of the process of use of new words in writing that would enable us to decide between a power law or a logarithmic description of the type-token ratio as a function of the number of text words.

2.4 The type-token ratio as a slowly varying function

Closer inspection of the function $\rho(n)n^d$, where the power law $\frac{A}{n^d}$ is obtained from a linear least-squares fit to a plot of $log(\rho(n))$ versus $log(n)$, does not yield a constant but typically an ultra-slowly varying function of n:

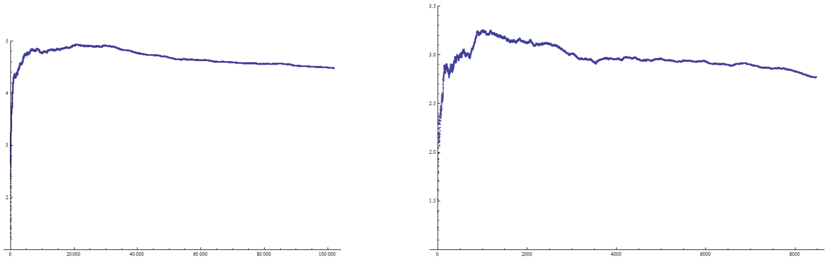

Figure 1. Plots of $\rho(n)n^d$ versus n where $\frac{A}{n^d}$ is the linear least squares fit to a plot of $\rho(n)$ versus n: Dante's *Divine Comedy* (left) and Einstein's *Sidelights on Relativity* (right). Note how the graphs increase relatively quickly initially and then decrease very slowly as n increases.

Typically, the functions $\rho(n)n^d$ where $\frac{A}{n^d}$ is the linear least squares fit to a plot of $log(\rho(n))$ versus $log(n)$ are very slowly decreasing after an initial rise. The initial rise of $\rho(n)n^d$ is typically logarithmic, a point we return to in section 2.3. The running variance of $\rho(n)n^d$ decreases as a power law with increasing n:

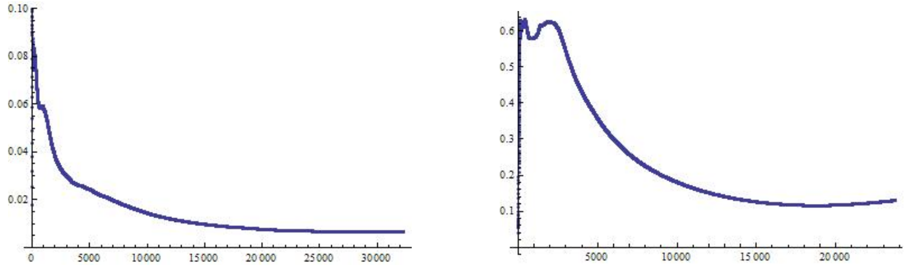

Figure 2. Running variance of $\rho(n)n^d$ where $\frac{A}{n^d}$ is the linear least squares fit to a plot of $log(\rho(n))$ versus $log(n)$: Dante's *Inferno* (left: approximated by a power law for $n \geq 1000$; $r^2 = 0.983$) and Beowulf (right: approximated by a power law for $n \geq 3000$; $r^2 = 0.946$).

The running variance of the functions $\rho(n)n^d$, where $\frac{A}{n^d}$ is the linear least squares fit to a plot of $log(\rho(n))$ versus $log(n)$, give a clear illustration of how a power law is not, overall, the best fit to the type-token ratio $\rho(n)$: if a power law held to a high degree of accuracy then $\rho(n)n^d$ would be very nearly constant and the running variances would be essentially 0, rather than a function that, after an initial rise, steadily decreases as a power function of n. Importantly, we find a relatively high degree of variance for small n, when the type-token ratio $\rho(n)$ is close to 1 - that is, when a majority of the words of the text to that point are new words.

2.5 Identifying a heavy tail of the text

To identify more accurately the later part of a text in which a power law seems to be a best fit to the type-token ratio we define $r^2(n_0)$ to be the linear least squares regression of $log(\rho(n))$ on $log(n)$ for $n \geq n_0$. We plot $r^2(n_0)$ versus n_0 and look for the first local maximum in this plot: at this value of n_0 we obtain the locally highest r^2 for a power law fit to the type-token ratio for $n \geq n_0$.

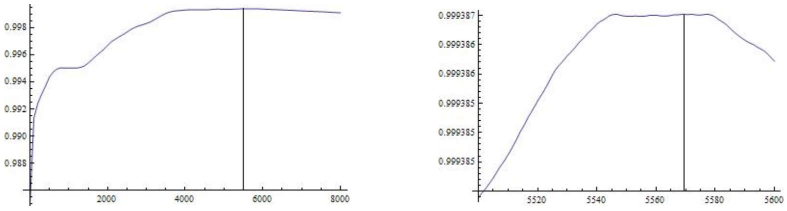

Figure 3. Left: A plot of $r^2(n_0)$ versus n_0 for Dante's *Inferno.* A local maximum appears around $n_0 = 5500$. Right: Restricting the plot to $5500 \leq n_0 \leq 5600$ locates a local maximum for $r^2(n_0)$ of 0.999 at $n_0 = 5569$.

The example, above, of a local maximum correlation coefficient for a linear regression of $log(\rho(n))$ on $log(n)$ shows how the plot of the (square of the) correlation coefficient versus the point in the text from which the correlation is done, provides an estimate of an initial segment of the text for which the correlation is not so high, and for which, from that point on is (locally) a maximum. For the example of Dante's *Inferno*, given above, a power law holds for $n_0 \geq 5569$ with the high r^2 value of 0.999. The first local maximum in the plot of $r^2(n_0)$ versus n_0 is generally not a global maximum. For example, for P.G. Wodehouse's *Right Ho Jeeves*, there are two further local maxima beyond the first local maximum:

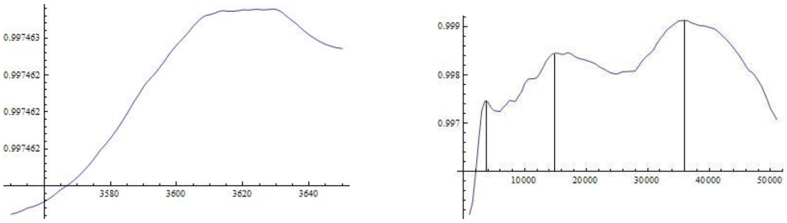

Figure 4. Left: the first local maximum, $r^2(n_0) = 0.997$, occurs at $n_0 = 3629$. Two further local maxima at $n_0 = 15138$ and $n_0 = 36027$ give slightly higher values for $r^2(n_0)$: 0.998 and 0.999 respectively.

2.6 The type-token ratio before the turnover point

We refer to the value of n_0 at which $r^2(n_0)$ first obtains a relatively high (typically ≥ 0.98) local maximum, as the *turnover point* of the text, and we denote this value of n_0 by n^*. The question we address here is the behavior of the type-token ratio prior to the turnover point n^*.

A regression analysis indicates that $\rho(n)n^d$ commonly increases approximately linearly or logarithmically for $n \leq n^*$, where d is the slope of a log-log plot of $\rho(n)$ versus n over the complete text.

2.6.1 Ratio of a linear and power function

When we suspect the type-token ratio is of the form $\rho(n) = \frac{A+Bn}{n^d}$ the question that arises is the optimum value of d for which $\rho(n)n^d$ increases linearly for $n \leq n^*$. To address this question we look at how the (square of the) correlation coefficient $r^2(d)$ for a regression of $\rho(n)n^d$ on n, for $n \leq n^*$, varies with d, and we choose the value of d that gives a maximum for the corresponding correlation coefficient. Consider, as an example, D.H. Lawrence's *England My England*, where the turnover point occurred very early, at $n = 212$, and prior to the turnover point $\rho(n) = \frac{1.194+0.844n}{n^{1.06}}$ with $r^2 = 0.998$. We see a much better fit to the type-token ratio prior to the turnover point than is given by the original power law, and we additionally obtain a power law $\rho(n) = \frac{4.092}{n^{0.323}}$ with $r^2 = 0.992$ for $n \geq 212$:

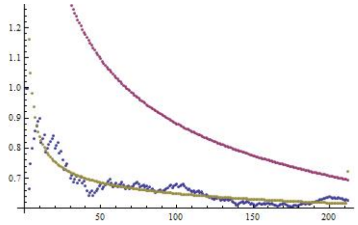

Figure 5. Comparison of $\rho(n)$ with the original power law for the entire text (top) and the modified power law for $1 \leq n \leq 212$ [D.H. Lawrence's *England My England*. Note that *England My England* is a collection of short stories written between 1913 and 1921].

When the turnover point n^* is on the order of several thousand words, the least squares estimate of $\rho(n)$ as $\frac{A+Bn}{n^d}$ for $n \leq n^*$ is generally poor for the first few hundred words, but very good over the range $1 \leq n \leq n^*$:

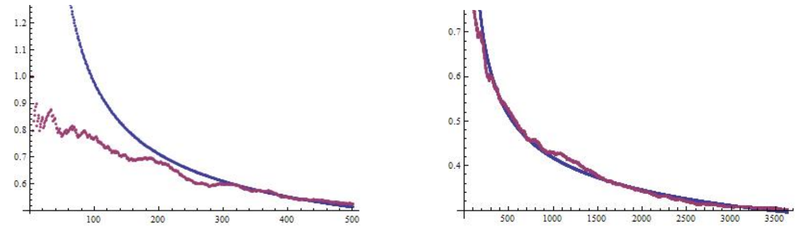

Figure 6. Left: Relatively poor approximation of $\rho(n)$ (bottom)for $n < 200$ by a least squares fit of $\frac{A+Bn}{n^d}$ (top). Right: good overall fit ($r^2 = 0.999$) of $\frac{A+Bn}{n^d}$ to $\rho(n)$ for $1 \leq n < n^*$: text, P.G. Wodehouse's *Right Ho Jeeves*.

2.6.2 Ratio of a logarithmic and power function

We have seen for a turnover point n^* on the order of several thousand, an estimate of an index d and a linear least squares fit to $\rho(n)n^d$ can lead to an excellent overall fit of a ratio of a linear and a power function to the type-token ratio up to the turnover point, but a poor fit at the beginning of the text. We adopted an alternative approach and used $Mathematica^{TM}$'s built-in **FindFit** command to find good fits to $\rho(n)$ for $1 \leq n \leq n^*$ of the form $\frac{A+Bn}{n^d}$ and of the form $\frac{A+Blog(n)}{n^d}$. Generally, **FindFit** did not produce a good fit of the form $\frac{A+Bn}{n^d}$ but did find good fits of the form $\frac{A+Blog(n)}{n^d}$. Shown below is the $Mathematica^{TM}$ fit of the form $\frac{A+Blog(n)}{n^d}$ to the type-token ratio up to the turnover point for the text P.G. Wodehouse's *Right Ho Jeeves* analyzed in 2.6.1 above:

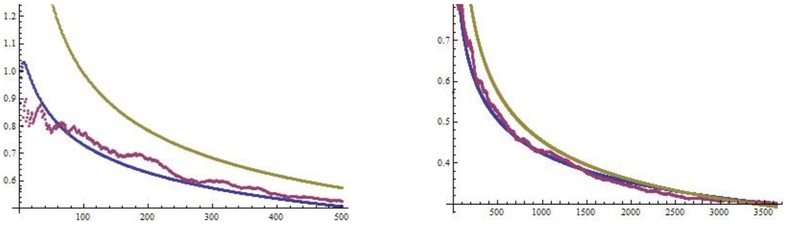

Figure 7. Plot of $Mathematica^{TM}$'s **FindFit** fit to $\rho(n)$ of the form $\frac{A+Blog(n)}{n^d}$ with $A = 0.572, B = 0.850, d = 0.394$. The original power law estimate from a log-log regression of $\rho(n)$ on n is shown at the top. Left: the plot for $1 \le n \le 500$. Right: The plot for $1 \le n \le n^* = 3629$.

Mathematica^{TM}'s **FindFit** fit to $\rho(n)$ of the form $\frac{A+Blog(n)}{n^{d_1}}$ for $1 \leq n \leq n^*$

Author - Text	# words	n*	A	B	d_1	$\rho(n) = A/n^d$ for $n \geq n^*$ (r^2)
Kant, I. - Kritik der reinen Vernunft	174877	837	1.099	0.209	0.227	$15.307/n^{0.473}(0.993)$
Darwin, C. - Origin of Species	149133	9583	0.623	0.885	0.408	$21.619/n^{0.509}(0.997)$
Milton, J. - Paradise Lost	80007	5716	0.807	0.5710	0.321	$12.605/n^{0.405}(0.995)$
Lawrence, D.H. - England My England	64108	221	0.974	0.055	0.136	$8.809/n^{0.382}(0.997)$
Balzac - Contes Bruns	60775	696	0.986	0.309	0.252	$4.238/n^{0.293}(0.989)$
Wodehouse. P.G. - Right Ho Jeeves	52832	3629	0.572	0.850	0.394	$4.735/n^{0.340}(0.997)$
Dante Alighieri - Inferno	32265	6054	0.854	0.561	0.322	$4.394/n^{0.292}(0.999)$
Voynich manuscript	31851	9408	0.737	0.510	0.275	$7.046/n^{0.299}(0.988)$
Traditional - Beowulf	23732	1812	0.754	0.460	0.301	$11.157/n^{0.427}(0.995)$
Wittgenstein, L. - Tractatus	22580	347	0.780	0.960	0.441	$5.300/n^{0.400}(0.996)$
Aaronsohn, A. - With the Turks in Palestine	17537	4293	0.891	0.591	0.337	$8.809/n^{0.382}(0.997)$
Einstein, A. - Sidelights on Relativity	8472	867	0.973	0.498	0.349	$5.027/n^{0.362}(0.998)$
Greer, G. - Worlds Apart	3557	1284	0.897	0.548	0.322	$3.881/n^{0.290}(0.997)$
Elliot, T.S.- Waste Land	3109	57	0.948	0.392	0.262	$2.489/n^{0.233}(0.970)$
McKenzie, N. - Master networker	2266	301	0.891	0.596	0.339	$2.476/n^{0.238}(0.990)$
Traditional Maori - Hinemoa	2005	247	0.886	0.710	0.437	$24.140/n^{0.653}(0.990)$

Table 3. *Mathematica*^{TM}'s **FindFit**fit to $\rho(n)$ of the form $\frac{A+Blog(n)}{n^{d_1}}$ for $1 \leq n \leq n^*$. Right hand column: generally high r^2 for a power law fit to $\rho(n)$ as a function of n for $n \geq n^*$

2.7 Self-organized criticality

The phenomenon of a power law decay for the type-token ratio with a high (≥ 0.98) and locally maximum r^2 after a certain point of the text, but not initially, is suggestive of the development of self-organized criticality ([16], [17]) because when the type-token ratio is well-described by a power law the number of runs of previously used words of length k or greater also decays approximately as a power law with k. In Bak's "avalanche" model of self-organized criticality [17], the runs of previously used words are the "avalanches". In fact, let us make the definition that an *avalanche* in a text is a run of previously used words. At the beginning of a text we have very small avalanches, but as the text progresses and new word use becomes more rare, according to a power law distribution, larger avalanches become more common, and the number of them of size $\geq k$ behaves approximately as a power function of k. It seems plausible that for the early part of a text, as identified by the turnover point, an author introduces new words into the text from an existing vocabulary until a point equivalent to a phase transition is reached, at which point the text becomes scale-invariant in terms of the behavior of the type-token ratio and runs of previously used words.

2.8 Relatively new words

For a positive integer k we denote by $t(n;k)$ the number of words that occur k or fewer times in the first n words of a text, and we set $\rho(n;k) := \frac{t(n;k)}{n}$. When $k = 1, \rho(n;k) = \rho(n)$ is the type-token ratio for the first n words. We found that for small $k, \rho(n;k)$ is of the form $\frac{L(n)}{n^d}$ where $L(n)$ is an ultra-slowly varying function. In other words, the same power law phenomena holds for the average number of relatively new words as holds for new words of an initial text segment, albeit with decreased correlation coefficients. Typically, r^2 for a log-log regression of $\rho(n;k)$ on n decreases approximately linearly with k:

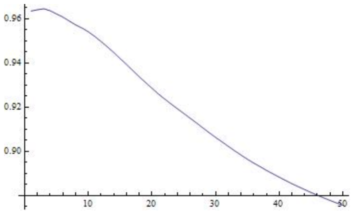

Figure 8. Typical linear decay ($r^2 = 0.990$) of r^2 for a log-log regression of $\rho(n;k)$ on n : text Aarohnson's *With the Turks in Palestine*.

3 Running variance of new words

The type-token ratio for the first n words of a text is the running average of the number of new words, up to that point in the text, and the type-token ratio is well-described by a power law decay after an initial logarithmic/power-law decay. We found that typically the running variance $v(n)$, of the number of new words up to the n^{th} word of the text, increases before decreasing logarithmically, linearly, or piecewise linearly, (or a combination over different segments of the text). An eventual logarithmic decrease of the running variance of the number of new words occurs, for example, in P.G. Wodehouse's *Right Ho Jeeves*:

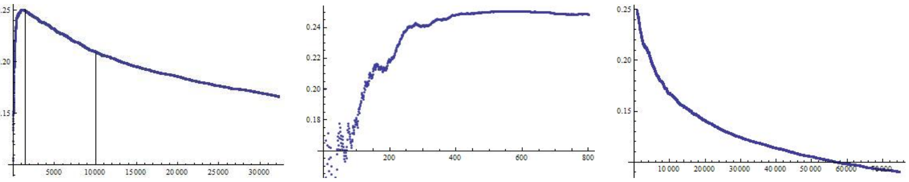

Figure 9. Left: Running variance $v(n)$ versus n for P.G. Wodehouse's *Right Ho Jeeves*. Middle: $v(n)$ initially increases, approximately logarithmically for $n \leq 800; r^2 = 0.880$. Right: $v(n) = 0.521 - 0.038 log(n)$ for $n \geq 801; r^2 = 0.998$.

An eventual linear decrease of the running variance of the number of new words occurs with Aaronsohn's *With the Turks in Palestine*:

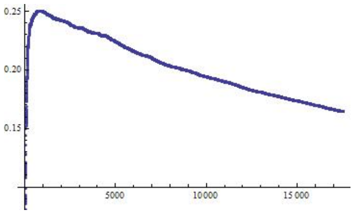

Figure 10. Running variance $v(n)$ versus n for Aaronsohn's *With the Turks in Palestine*. $v(n) = 0.249 - 5.19 \times 10^{-6} n$ for $n \geq 415; r^2 = 0.986$. The slope of the linear part of the graph is nearly 0, so the variance is practically constant from that point on.

Einstein's *Sidelights on Relativity* yields two regions where the decrease in the running variance $v(n)$ is best described by different linear segments:

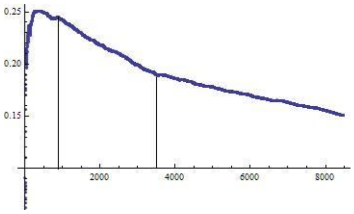

Figure 11. Running variance for Einstein's *Sidelights on Relativity*. For $n \leq 1400, v(n)$ increases and then decreases slightly. For $1400 \leq n \leq 3500, v(n) = 0.260 - 2.04 \times 10^{-5} n, r^2 = 0.998$, and for $n \geq 3500, v(n) = 0.217 - 7.72 \times 10^{-6} n, r^2 = 0.997$.

The point around $n = 3500$ where the running variance changes from one linear piece to another, corresponds almost exactly to the point at which Einstein begins a new section of his text entitled *"GEOMETRY AND EXPERIENCE. An expanded form of an Address to the Prussian Academy of Sciences in Berlin on January 27th, 1921."* It is not unreasonable to infer from the change in the behavior of the running variance of the new words, and the beginning of this new section of his work at almost exactly the same point, that Einstein was writing in two different modes before and after 3500 words of the text of *Sidelights on Relativity.*

Finally, a combination of eventual linear and then logarithmic decrease occurs with Dante's *Inferno*:

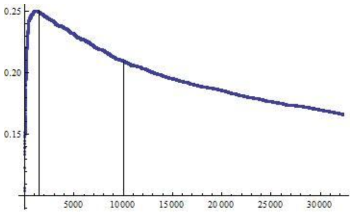

Figure 12. Running variance for Dante's *Inferno.* For $n \leq 1400, v(n)$ increases then decreases linearly $(v(n) = 0.255 - 4.680 \times 10^{-6}n, r^2 = 0.998)$ until approximately $n = 10000$ at which point $v(n)$ decreases logarithmically:$v(n) = 0.545 - 0.036log(n), r^2 = 0.999$.

The point at which the running variance changes from linear to logarithmic decay corresponds approximately to the beginning of Canto XII of the *Inferno.*

4 Entropy

The i^{th} word $w_i \in T_n$ has a relative frequency of occurrence $\omega_{n,i}$ in T_n. Given T_n, we regard $\omega_{n,i}$ as the probability of occurrence of the word $w_i \in T_n$. For this probability distribution on T_n the Shannon entropy of the initial segment of text T_n is $H(n) = -\sum_{i=1}^{n} \omega_{n,i} log_2(\omega_{n,i})$. This amounts to treating each T_n as a self-contained text, for statistical purposes, with the point of examining how the entropy changes as T_n is enlarged to T_{n+1} by the addition of a new word or a previously used word. We examined the variation of $H(n)$ with n for a variety of literary texts. First, some general observations are in order. The addition of a new word, in T_{n+1} but not in T_n, necessarily increases the entropy - that is, $H(n+1) > H(n)$ when T_{n+1} is formed by the addition of a new word to T_n . This is because, as one easily sees,

$$H(n+1) = \frac{n}{n+1}(H(n) - log_2(\frac{n}{n+1}) + \frac{1}{n+1}log_2(n+1))$$

so

$$H(n+1) - H(n) = \frac{1}{n+1}(-H(n) - nlog_2(n) + (n+1)log_2(n+1)) \geq$$

$$\frac{1}{n+1}(-log_2(n) - nlog_2(n) + (n+1)log_2(n+1)) = log_2(n+1) - log_2(n) > 0$$

since $H(n) \leq log_2(n)$.

The effect on the entropy of a previously used word is a little more subtle. If $w_n \in T_{n-1}$ then $w_n = w_j$ for some $1 \leq j \leq n-1$ so $H(n) =$

$$-\sum_{i \neq j}^{n} \frac{\omega_{n-1,i}(n-1)}{n} log_2(\frac{\omega_{n-1,i}(n-1)}{n}) - \frac{\omega_{n-1,j}(n-1)+1}{n} log_2(\frac{\omega_{n-1,j}(n-1)+1}{n})$$

and from this we see that $n[H(n-1) - H(n)] = H(n) + (n-1)log_2(\frac{n-1}{n})$

$$-\frac{\omega_{n-1,j}(n-1)}{n} log_2(\frac{\omega_{n-1,j}(n-1)}{n}) + \frac{\omega_{n-1,j}(n-1)+1}{n} log_2(\frac{\omega_{n-1,j}(n-1)+1}{n})$$

so for $n >> 1$ we have

$$n[H(n-1) - H(n)] \approx H(n-1) - log_2(n) - \frac{1}{log(2)} + log_2(\omega_{n-1,j}(n-1)+1).$$

Therefore, $H(n) < H(n-1)$ when $log_2(\omega_{n-1,j}(n-1)+1) > \frac{1}{log(2)} + log_2(n) - H(n-1)$, that is, when $\omega_{n-1,j}(n-1) > 2^{\frac{1}{log(2)}} n 2^{-H(n-1)} - 1 = en2^{-H(n-1)} - 1$. So typically, if $n = 1000$ and $H(n-1) \approx 8$ then the number of occurrences of the previously used word w_n in $T(n-1)$ should be at least 10 in order that $H(n) < H(n-1)$. The moral is that if a previously used word is relatively rare then the entropy rises as the word is used again, but at a reasonable rate of occurrence the entropy falls. Thus, as words are used progressively more often the rate of increase of the entropy is dramatically slowed.

Empirically, we find that the entropy $H(n)$ of T_n increases logarithmically with n. This would follow if it were true that $\left|H(n+1) - H(n)\right| \approx \frac{B}{n}$, for some constant $B > 0$, when $n >> 1$. However, empirically, the latter does not seem to be the case. For example, a plot of $H(n)$versus n for Dante's *La Divina Comedia* yields an approximate logarithmic increase ($r^2 = 0.899; p \approx 0$) but a plot of $n(H(n+1) - H(n))$versus n is neither constant, nor ultra-slowly varying (Figure 1):

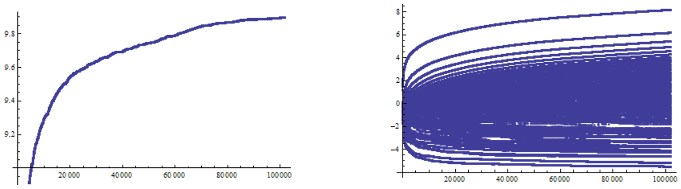

Figure 13. Figure 1. $H(n)$ versus n, and $n(H(n+1) - H(n))$versus n

The entropy differences $H(n+1) - H(n)$ are reasonably large for small n, with a predominant $\frac{1}{n}$-decay, but very small and more scattered for $n >> 1$:

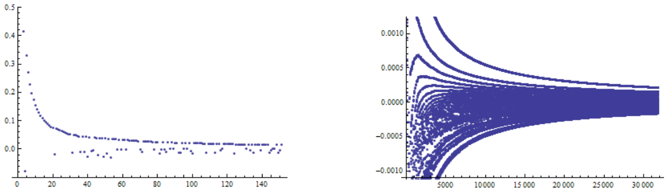

Figure 14. Entropy difference for Dante's *Divine Comedy.* Small n (left) and large n (right)

The entropy increase with the length of text was well-fitted by a logarithmic function for all the texts we examined. However, certain texts contain blocks that show significant deviation from a logarithmic increase. One such example is Einstein's *Sidelights on Relativity*, for which there are small, but identifiable, blocks of text where the running entropy decreases:

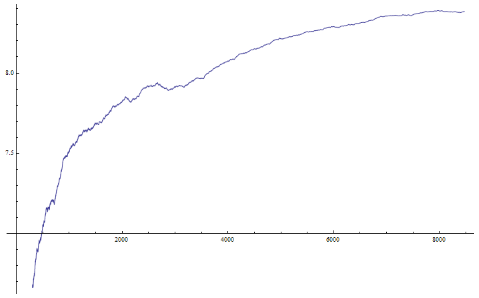

Figure 15. $H(n)$ versus n for Einstein's *Sidelights on Relativity.* There are small, but significant, decreases in the running entropy around 2000 and 2600 words.

A more significant example occurs in the Voynich manuscript. This is MS 408 of the Beinecke library at Yale University. It is a still mysterious, undeciphered manuscript written using unusual symbolic forms, but apparently representing a text with linguistic structure [15]. Using the Takahashi transcription [16] of these symbolic forms we plotted the entropy $H(n)$ of the first n words of the Voynich text as a function of n. As for all the other texts we examined, $H(n)$ varies approximately logarithmically with n ($r^2 = 0.926; p \approx 0$). However, there is a significantly large bock of the Voynich text, about 5000 words from the first 12000 words of the text on - approximately 16% of the total text - for which the entropy decreases. This necessarily indicates a large degree of repetition of words that have been used significantly often in the text before this point.

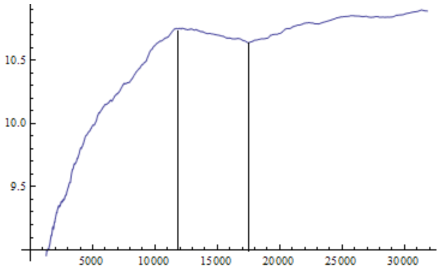

Figure 16. Figure 3. $H(n)$ versus n for the Voynich manuscript. Note the decrease in entropy from $n \approx 1200$ to $n \approx 17000$

For the Voynich text, the rate of increase of entropy for the first 12,000 words is markedly different to that for the last 14,000 words. This indicates not only a greater rate of use of new word types in the first 12,000 words than in the last 14,000 - a fact not unusual in itself - but also a possibly different *model* of word use in these two blocks of text.

5 Concluding remarks

We have shown that the running average of new words in literary texts behaves initially as a ratio of a logarithmic function to a power function, but eventually settles down to a pure power function. This phenomenon is true over a range of text lengths, genres and languages. We found that the transition point from the initial behavior to power law decay was quite variable, ranging from a few tens of words, to tens of thousands of words. After the transition, the text exhibits statistical properties of self-organized criticality, which are not evident in the initial phase. This prompts us to believe that the initial phase is one in which a writer uses new words to begin to build a coherent, if implicit, model of the text. It is possible that something akin to a phase transition occurs after that point.

We found that the running variance of new word types does not decay as a power function but is typically linear, logarithmic, or a piecewise combination of the two. The plot of the running variance seems to have some value as a diagnostic tool in ascertaining if parts of a text are produced in different writing modes - Einstein's *Sidelights on Relativity* being a striking example.

The running entropy generally increases logarithmically with the length of an initial segment of text. This finding has implications for the use of new or relatively rare words. If an author were to reach a point in a text where they were now using only words that had already been commonly used in the text the entropy would show a marked decline. With one or two notable exceptions - the Voynich manuscript and small segments of Einstein's *Sidelights on Relativity* - we do not see this happen: the overall increase of the running entropy is logarithmic. The implication of this fact is that authors of literary texts write with continuing novelty in the use of new or relatively rare words, no matter how long the text. The fact that the running average of new words eventually decreases as a power law of the length of text, tells us that, and how, new word use becomes rarer as the text progresses. The results we found on the use of relatively rare words, those used only a few times to a given point of the text, tells us that the running average of such words also decreases eventually as a power function of the length of text. Therefore, the overall logarithmic rise in the entropy is due to the sustained use of relatively novel words. The empirical evidence is that a writer of a literary text simply cannot write arbitrarily long segments of text with words that they have already used with high frequency in that text.

References

[1] G. K. Zipf, The Psychobiology of Language, George Routledge & Sons, 1936.

[2] G. K. Zipf, Human Behaviour and the Principle of Least Effort, Addison-Wesley, 1949.

[3] B. Mandelbrot, Information Theory: Third London Symposium, (1956).

[4] B. Mandelbrot, Information theory and psycholinguistics. In Wolman B. B. and E. N. Nagel (Eds.) Scientific Psychology: Principles and Approaches. New York: Basic Books. pp. 550-562 (1965)

[5] A. F. Parker-Rhodes, T. Joyce, Nature, 178, 1308 (1956).

[6] A. S. Corbet, R. A. Fisher, C. B.Williams, J. Anim. Ecol., 12, 42 (1943).

[7] I. J. Good, Biometrika, 40, 237 (1953).

[8] C. E. Shannon, Bell System Tech. J., 27, 379 and 623 (1948).

[9] N. Wiener, Cybernetics, MIT-Wiley, 1949.

[4] I. J. Good, Probability and the Weighing of Evidence, 1950.

[11] G. Herdan, Language as Choice and Chance, Noordhoff, 1956.

[12] G. Herdan, The Advanced Theory of Language as Choice and Chance. Springer, 1966.

[13] L. L. Goncalves, L. B. Goncalves, Fractal power laws in literary English. Physica A, 360 (2), 557-575 (2006).

[5] S. I. Resnick, Heavy-Tail Phenomena, Springer, 2007.

[6] G. Landini, Evidence of linguistic structure in the Voynich manuscript using spectral analysis, Cryptologia, (2001).

[16] P. Bak, C. Tang and K. Wiesenfeld, Self-organized criticality, Physical Review A 38: 364–374, (1988).

[17] P.Bak, How Nature Works: The Science of Self-Organized Criticality, Springer-Verlag, 1999.

[18] T. Takahashi, The Most Mysterious Manuscript in the World, www.voynich.com, 1999.